DOCUMENT RESUME

ED 341 713 TM 017 898

AUTHOR Cesare, Steven J.; And Others
TITLE A Predictive Validation Study of the Methods Used To Select Eligibility Technicians.
PUB DATE Jun 91
NOTE 48p.; Paper presented at the Annual Meeting of the International Personnel Management Association Assessment Council (15th, Chicago, IL, June 23-27, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Affirmative Action; *Employment Interviews; Ethnic Groups; Evaluation Methods; *Government Employees; *Job Performance; Local Government; Multiple Choice Tests; Occupational Tests; Personnel Evaluation; *Personnel Selection; *Predictive Validity; Program Evaluation; *Test Bias
IDENTIFIERS California (San Diego County); *Eligibility Technicians

ABSTRACT

The effectiveness of the current selection system used to hire eligibility technicians was studied, and the predictive relationship between existing selection instruments and job performance criteria was investigated. Eligibility technicians are employees of San Diego County (California) whose duties include determining the eligibility of applicants for social services. The current selection system consists of a multiple-choice examination and a composite assessment including a training and experience application blank and an interview. Job performance criteria were subjective (supervisory ratings and factor scores) and objective (job tenure and turnover). Scores on the subjective criteria were taken at three points in time: mid-probation, final probation, and first annual performance review. Information was gathered for all applicants for the position of eligibility technician between 1986 and 1990 (449 employees at mid-probation, 409 at final probation, and 250 at the first annual performance review). Applicants included Blacks, Hispanic Americans, Asian Americans, American Indians, Whites, and Filipinos. The written test was a valid predictor of job performance, with a positive relationship between the test and supervisory ratings at all times. The application and interview did not predict job performance. The written test had an inverse relationship with tenure and was significantly correlated with turnover. The written test was a fair and unbiased estimate of job performance that did not discriminate against applicant classes protected by the county. Suggestions for improvement of the selection process are given. Thirty-six tables and a list of 101 references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEVEN J. CESARE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED341713

A Predictive Validation Study of the Methods Used to Select Eligibility Technicians

Steven J. Cesare
Department of Human Resources
Recruitment and Assessment Division

Mark H. Blankenship
Department of Social Services
Management Services Bureau

Patrick W. Giannetto
Department of Human Resources
Recruitment and Assessment Division

Mark Z. Mandel
Department of Human Resources
Recruitment and Assessment Division

Running head: ET VALIDATION

This paper was part of the symposium entitled, *A psychometric and fiscal assessment for selecting eligibility workers: Multiple agency results* presented at the fifteenth annual conference of the International Personnel Management Association Assessment Council, Chicago, Illinois, June 1991.

TM017988

BEST COPY AVAILABLE

Abstract

The intent of this report was to assess the effectiveness of the current selection system used to hire Eligibility Technicians and to investigate the predictive relationship between the existing selection instruments and job performance criteria. The current selection system consists of a written multiple-choice test and a composite instrument inclusive of a training and experience application blank and a structured interview. The job performance criteria for this project were of two types: subjective and objective. The subjective criteria were supervisory ratings and derived factor scores taken at three points in time: mid-probation, final-probation, and first-annual performance review. The objective criteria were job tenure and turnover. Selection system data were gathered on all applicants for the position of Eligibility Technician between the years of 1986 and 1990. Criteria information was collected for all applicants who were offered employment and accepted a position with the County of San Diego.

The results indicated a consistent pattern across a majority of the criterion measures. (1) The written test was found to be a valid predictor of subsequent job performance in that the applicants who scored well on the written test received higher job performance ratings than those applicants who scored low on the test. (2) The positive relationship between the written test and supervisory ratings was evident for the mid-probation, final-probation, and first-annual performance reviews. (3) The application/interview composite failed to predict job performance for each criterion. (4) The written test had an inverse relationship with tenure and was significantly correlated with turnover. (5) It was statistically determined that the written test was a fair and unbiased estimate of job performance; it did not discriminate unfairly against any of the protected classes identified by the County of San Diego.

Interpretative suggestions were provided on the three key areas: selection instruments, criteria, and validation design. Despite the positive results concerning the written test, suggestions for improving the psychometric and practical aspects of the selection system were provided. With regard to criterion issues, several interventions were put forth: a diverse complement of multiple objective and subjective criteria, multiple rating sources, and rater training designed to minimize judgmental bias. As an interim strategy, it was mentioned that a quantitative index of content validity should be adopted; a discussion of indirect validation methods ia also presented.

A Predictive Validation Study of the Methods Used to Select Eligibility Technicians

A sound and optimal human resource system can be reflected in multidimensional terms: functional accountability (i.e., judicious use of tax payer funds), efficient productivity (i.e., increasing the ratio between performance and employee), and equitable utilization (i.e., minimizing human resources costs). The fundamental premise for this system is psychometric validity. Demonstration of validity implies productivity (Schmidt & Hunter, 1980), cost effectiveness (Schmidt, Hunter, McKenzie, & Muldrow, 1979), and legal compliance (Kleiman & Faley, 1985).

In reference to employee selection, validity refers to the inferences that can be made from the scores on a selection test; validity is not an inherent quality of the test itself. These inferences are limited by specific parameters: the test itself, the criterion measure, the test takers, and the psychometric properties associated with each (e.g., reliability, fidelity, and homogeneity, respectively). Any alteration of these factors necessarily impacts the validity of the inferences. Case in point, the popular phrase "the test is valid" incorrectly implies that the test is the sole determinant of validity and assumes that because the test is currently "valid" it will always be valid in the future as well.

Job Analysis

As suggested by the *Uniform Guidelines on Employee Selection Procedures* (1978), a validation project should be premised on a job analysis of the job class under investigation. The job analysis serves many vital functions: (a) determination of the salient knowledges, skills, and abilities inherent within the job; (b) identification of the appropriate criteria needed to assess job performance, systematically and accurately; and (c) illustration of the methodological rigor involved which has been frequently scrutinized in litigation (cases include: Association Against Discrimination v. City of Bridgeport, 1978; Rivera v. City of Wichita Falls, 1982; Walls v. Mississippi Department of Public Welfare, 1982). In their review of this area, Thompson and Thompson (1982) summarize the requisite steps for a legally defensible job analysis. The job analysis data used for the current class of Eligibility Technician represent the logical foundation for the three predictive instruments: a written multiple-choice test, a training and experience application blank, and an interview.

Predictor Issues

Written Test. Title VII of the 1964 Civil Rights Act forbids the use of employment tests that are discriminatory unless the employer can demonstrate that the test has a manifest relationship to job performance (i.e., validity). The landmark case of Griggs v. Duke Power Company (1971) crystallized that point by confirming that written tests which do not measure job performance may be discriminatory.

The *Uniform Guidelines* (1978) state that whether a selection test is standardized (e.g., Wonderlic Personnel Test, Bennett Mechanical Comprehension Test) or developed internally (e.g., job knowledge test) by an organization, it is incumbent upon the test user to conduct either an empirical (criterion-related or construct validity) or content validation study to demonstrate the job relatedness of the written test.

It is the intent of this report to validate the effectiveness of the selection predictors by using a criterion-related strategy. The degree of relationship between an employee's test score and subsequent job performance criteria will be measured statistically and viewed in context of the significance level specified in the *Uniform Guidelines*. According to the *Uniform Guidelines*, empirical validation is determined if a significant relationship exists between the selection test and the criterion variable; significance is equivalent to a probability level of .05 or less, which means that the relationship is sufficiently large as to have a probability of no more than one in twenty to have occurred by chance.

Application Blank. This type of selection instrument serves to (a) decide if an applicant meets the minimum requirements for a position and/or (b) assess the applicant's strengths and weaknesses for the job

position. Though limited, research documenting the accuracy of information on self-report application blanks does exist. Falsification of application blank data can range from mild exaggerations of one's college grade point average to complete lies about one's previous employers, job duties, and degrees held. To date, the research in this area suggests that application blanks possess some fraudulent data (Goldstein, 1974; Mosel & Cozan, 1952). Likewise, a recent meta-analysis has found that different scoring strategies produce a wide range of validity coefficients (McDaniel, Schmidt, & Hunter, 1988). While the research need in this area is obvious, the practitioner must remain cautious in interpreting an applicant's personal history of his/her training, education, and employment. Advisedly, application blanks which consist of questions that are objective, easily verified, indicate the presence of a lie scale, and/or use multiple, trained evaluators serve to improve the intention of the instrument (Ash & Levine, 1985; Cascio, 1987).

The Equal Employment Opportunity Commission (1981) has stated that:

> Employment application forms ... have traditionally been instruments for eliminating, at an early stage, 'unsuited' or 'unqualified' persons from consideration for employment and have been used in such a way as to restrict or deny employment opportunities for women and members of minority groups (p. 1).

As such, the application blank is held to the same legal standards set forth by the *Uniform Guidelines*. Questions found on the application blank must: (a) not screen out protected group members disproportionately, (b) be predictive of subsequent job performance, or (c) be justified in terms of business necessity. Clearly, the validity of the application blank must be demonstrated to ensure its purpose of predicting job relatedness.

Interview. The interview is one of the most popular selection methods used throughout industry (Harris, 1989; Ulrich & Trumbo, 1965). Indeed, Miner and Miner (1978) found that 56% of 200 companies surveyed in 1976 identified the interview as their most important selection device. More surprising, is that only 4 of those 200 companies had conducted a validation study involving the interview. The recent Supreme Court decision in the case of Watson v. Fort Worth Bank & Trust (1988) is likely to stimulate both additional investigation of the validity of selection interviews and improved validation efforts.

A common problem with the interview, which mitigates its validity, is the likelihood that subjective factors may produce biased decision making and possible discrimination concerning employment (Arvey, 1979; Schmitt, 1976). The legal literature is replete with cases where judges have addressed poorly developed interviews and interview content which is not job-related. Each of these factors invites subjective judgment and usually eventuates in judicial reprimand (EEOC v. Detroit Edison, 1975; King v. New Hampshire Department of Resources and Economic Development, 1977; Weiner v. County of Oakland, 1976).

Attempts to improve the interview include: developing a structured interview which asks only job-related questions; constructing interview questions that are situationally oriented; providing training based on sound learning principles (e.g., modeling, practice and feedback) to potential interviewers; use of a systematic scoring system; and having multiple interviewers evaluate the interviewee. An interview incorporating these qualities is likely to reduce biased decisions which, in turn, should increase its validity.

Traditional narrative literature reviews of interview validity have consistently concluded that the interview is a poor selection instrument (Arvey & Campion, 1982; Reilly & Chao, 1982; Schmitt, 1976). However, the advent of meta-analytic techniques has served as the foundation for a more quantitative review of the literature (McDaniel, Whetzel, Schmidt, Hunter, Maurer, & Russell, 1987).

In their review of criterion-related validity studies, McDaniel et al. (1987) found that the average validity coefficient was .22 for all interviews using job performance as the criterion; after correcting for unreliability in the criterion and range restriction, the coefficient jumped to .41. Furthermore, they found that structured interviews yielded higher validities than unstructured interviews; validity was higher when the criterion was job performance rather than training performance; and interviews conducted for research

purposes produced higher validity coefficients than those done for administrative decisions. More than simply illustrating the validity of the interview, their review demonstrates that methodological rigor and appropriate implementation (e.g., rating guide, multiple raters) can effect incremental outcomes (e.g., functional, statistical, and financial) that have direct benefit to an organization.

Criterion Issues

The *Uniform Guidelines* (1978) state that those criteria used in a validation study should be predicated on a thorough job analysis, thereby representing important or critical work behaviors or work outcomes. Despite this instruction, most criteria are usually developed in cursory fashion and as such, are either deficient or contaminated by other non-job related factors (Nathan & Alexander, 1988).

Periodic reviews of validation studies have consistently found performance ratings to be the most frequently used criterion measure (Guion, 1965; Landy & Trumbo, 1980; Schmitt, Gooding, Noe, & Kirsch, 1984). Despite this prevalence, the professional literature is filled with evidence identifying performance appraisal ratings as incomplete, biased, and inaccurate (Arvey & Faley, 1988; Barrett & Kernan, 1987; Feild & Holley, 1982; Kleiman & Durham, 1981).

Likewise, the legal literature has also highlighted the inadequacy of supervisory ratings of performance appraisal. The cases of Dickerson v. U.S. Steel Corporation (1978), Walls v. Mississippi State Department of Public Welfare (1982), and Association Against Discrimination v. City of Bridgeport (1978), each symbolize an instance in which a judge criticized the use of performance ratings as the criterion employed in a validation study.

A suitable complement to supervisory ratings is a more objective criterion measure. Objective criteria are more reliable measurements of job performance and tend to yield higher validity coefficients than performance ratings (Schmitt, et al., 1984) and when combined with a performance appraisal rating, the combined criteria tend to be more accurate description of job performance by accounting for an additional proportion of the overall "criterion space" (i.e., the complete scope of an employee's job performance). In substantiation of this, a recent meta-analysis by Heneman (1986) found a weak intercorrelation between performance ratings and results-oriented criteria (corrected mean validity coefficient of .27).

This project will incorporate the use length of service and employee turnover as objective criterion measurements. Schmitt, et al. (1984) found that the average validity coefficient with turnover as the criterion was .246. It is believed that this index will provide tangible insight into the relationship between the selection system and an unbiased outcome measure. In like manner, the criterion of length of service will demonstrate the degree of relationship with long-term employment; a criterion of immense value to several human resource strategies (e.g., planning, training, benefits).

Validation Strategy

Criterion-Related Validity. While the *Uniform Guidelines* state a preference for predictive validation studies, the judicial literature demonstrates that concurrent validation is the norm when predicting a job performance criterion (Kleiman & Faley, 1985). Judges have frequently criticized the use of concurrent validation; often citing the perspective that current job experience and homogeneous job incumbents represent biased test performance (Dickerson v. U.S. Steel Corporation, 1978; Walls v. Mississippi State Department of Public Welfare, 1982). Subsequent research in this area has shown that the biasing effects of those factors is less influential than originally thought and that there are minimal differences between the average validity coefficients produced by the two strategies: .34 for concurrent and .26 for predictive (Schmitt, et al., 1984).

Given the context of the current project and the recommendations put forth in the *Uniform Guidelines*, the present study will employ a predictive validation design. It is believed that reliance on this design is methodologically appropriate and in the event of litigation, illustrates the presence of minimal bias

(real or perceived) to the judge.

The present validation study will attempt to represent the ethnic and gender demographics of Eligibility Technicians in an equitable fashion. It is believed that by sampling all Eligibility Technicians between 1986 and 1990, an accurate ethnic and gender representation will result. This degree of accuracy provides results that are generalizable and interpretations that can be made with confidence.

Method

Definition of Job Class

According to the *Uniform Guidelines*, it is suggested that all job titles included within a given validation study be defined in terms of the corresponding titles and codes found in the Dictionary of Occupational Titles (1977). Following from that recommendation, the position under study in this project, Eligibility Technician, is interpreted as Eligibility Worker (# 195-267-010).

> This job class: Interviews applicants or recipients to determine eligibility for public assistance; interprets and explains rules and regulations governing eligibility and grants, methods of payment, and legal rights to applicant or recipient; records and evaluates personal and financial data obtained from applicant or recipient to determine initial or continuing eligibility, according to departmental directives; initiates procedures to grant, modify, deny, or terminate eligibility and grants for various aid programs, such as public welfare, employment, and medical assistance; authorizes amount of grants, based on determination of eligibility for amount of money payments, food stamps, medical care, or other general assistance; identifies need for social services, and makes referrals to various agencies and community resources available; prepares regular and special reports as required, and submits individual recommendations for consideration by supervisor; and prepares and keeps records of assigned cases.

Job Analysis

The job analysis inventory used in this study involved the participation of subject matter experts for the position of Eligibility Technicians. These experts were either experienced job incumbents or supervisors of incumbents and were thereby highly familiar with the duties and responsibilities involved in the job being studied.

The major reason for using an inventory approach to job analysis is to obtain information about the tasks and knowledge areas from a large number of job incumbents. This approach provides that an opportunity for suitable cross-sectional responses are given. The results from the job analysis indicated the existence of five job domains. These domains and a brief description for each are presented below.

> *Communication* - The ability to convey and receive information on a variety of topics (e.g., public assistance, notification, making referrals) verbally and in writing, to various parties (e.g., clients, collateral contacts, other agencies).
>
> *Budgeting and Determining Eligibility* - The ability to evaluate the client's personal and financial situation, refer to established criteria, and then draw appropriate conclusions in determining assistance.
>
> *Recording* - The ability to establish and maintain case files for information, departmental memos, and instructions; prepares activity reports; records time, keeps and submits logs.

Verifying - The ability to view and interpret a variety of forms and documents and to evaluate whether the documentation satisfies legal requirements.

Workload Management - The ability to: organize one's time, desk, and workload efficiently; meet agency and state expectations of volume and accuracy of work completed; and follow directions, policies, and procedures effectively.

Sample Characteristics

The total sample for this project was 3,614. Of this number, ethnicity data are available on 3,268. In terms of sample ethnicity, there were: 585 blacks (18%), 591 Hispanics (18%), 254 Asians (8%), 32 American Indians (1%), 1,517 whites (46%), and 289 Filipinos (9%). The gender composition of the sample was 2,315 females (71%) and 953 males (29%).

The total sample was used exclusively in conducting analyses on the evaluation of the selection test. Due to the fact that some applicants failed the written test, the sample size was reduced for subsequent analyses on the application blank and interview. The complete sample size for analyses conducted on the application blank was 1,313: 137 blacks (10%), 258 Hispanics (20%), 71 Asians (5%), 11 American Indians (1%), 792 whites (61%), 44 Filipinos (3%), and 919 (70%) were female. Demographic data were available for 1,303 interviewees: 135 blacks (10%), 260 Hispanics (20%), 70 Asians (5%), 11 American Indians (1%), 784 whites (61%), 43 Filipinos (3%), and 900 (69%) were female.

The precise sample size for the validation equation was contingent upon which of three validation samples was used (e.g., mid-probation, final-probation, or first-annual performance review). A summary of the demographic variables for each of the three validation samples are presented below.

Mid-Probation Validation Sample: The size of this validation sample was 449; demographic data were available for 439. Of this number there were: 36 blacks (8%), 110 Hispanics (25%), 24 Asians (7%), 4 American Indians (1%), 248 whites (56%), 15 Filipinos (3%), and 309 females (70%).

Final-Probation Validation Sample: The size of this validation sample was 409; demographic data were available on 403. For this sample there were: 31 blacks (7%), 101 Hispanics (25%), 24 Asians (6%), 4 American Indians (1%), 230 whites (57%), 13 Filipinos (3%), and 279 females (69%).

First-Annual Validation Sample: The size of this validation sample was 250. This sample consisted of: 13 blacks (5%), 64 Hispanics (25%), 16 Asians (6%), 3 American Indians (1%), 144 whites (57%), 10 Filipinos (4%), and 169 females (67%).

Description of Selection Procedures

Written Test. The primary selection instrument used in appointing Eligibility Technicians is a 106 item, written multiple choice test. The written test was divided into six subscales. The total from each subscale was aggregated equally to form an unweighted-unidimensional composite which serves as the test score for each job applicant. The Kuder-Richardson 20 coefficient for the entire test is .90. The subscales are presented below; the Kuder-Richardson 20 reliability coefficient for each subscale appears in parentheses.

Reading Comprehension (.69) - A test of the applicant's ability to read 1-2 short paragraphs and then answer a series of questions about the material presented.

Time Management and Workload Organization (.42) - A test designed to measure the applicant's ability to make decisions through efficient prioritizing of information, situational

logic, and resourcefulness.

Comprehending and Applying Eligibility Regulations (.74) - A test which measures the ability of the applicant to read and understand eligibility policies, and answer questions concerning those policies. Some questions require the applicant to perform basic arithmetic operations.

Computational Skills (.74) - A test which measures basic mathematical ability (i.e., addition, subtraction, multiplication, and division) as well as simple arithmetic reasoning.

Interviewing (.65) - A test which assesses the applicant's understanding of interviewing dynamics, the ability to gather information effectively, and provide an empathic and comfortable environment for the interviewee.

Interpersonal Relations (.78) - A test of the applicant's ability to communicate effectively, act professionally, and facilitate an attitude of teamwork with subordinates, peers, and/or supervisors.

Application Blank. The application blank consisted of four questions each relating to a specific content area: rules and regulations, managing large case loads, interviewing, and automation. The first three questions were assessed using a four-point rating scale (i.e., 0-1 Unsatisfactory; 2 Average; 3 Satisfactory), while the question on automation was measured dichotomously (i.e., 0 or 1). The item scores were summed into a composite which represented the applicant's final application rating. Each application blank was reviewed and assessed by two raters. The average inter-rater reliability was found to be .97.

Interview. The interview protocol consisted of five evaluation factors: four were questions asked to the applicant, while the fifth factor was an assessment of the interviewee's communication skills. Each of the five factors was measured using a six-point rating scale (i.e., 0-1 Unsatisfactory; 2-3, Average; 4-5, Satisfactory). The item scores were summed into a composite which represented the applicant's final interview rating. Each applicant was interviewed and assessed by two raters. The average inter-rater reliability for this technique was .96.

Though identified as a selection system, the three predictors were technically used as screening devices. The singular intent of the predictors was to distinguish between those candidates who demonstrate a satisfactory knowledge of the minimum qualifications necessary for suitable performance as an Eligibility Technician and those candidates who do not demonstrate that competency. It is an applicant's composite score on the three predictors which determined his/her ordinal position on the Eligibility list. Actual hiring decisions were the result of departmental interviews, conducted by the Department of Social Services which employs Eligibility Technicians.

Criterion Measures

Subjective Criterion Measures. The performance appraisal instrument used in this validation study was a standardized 22 item, 5-point graphic rating scale; standardized to allow for application to all classified employees throughout the entire County. This rating instrument contained 21 itemized performance dimensions and one global overall assessment rating. The five anchors for each of the 22 ratings were: (5) Outstanding, (4) Above Standard, (3) Standard, (2) Improvement Needed, and (1) Unsatisfactory. Supervisors were also given an option to indicate "Does Not Apply" to each of the 21 dimension items.

For the current project, performance appraisal ratings were culled from the personnel files of all Eligibility Technicians employed between 1986-1990. Three sets (the 21 individual job component ratings and the overall assessment rating) of performance ratings were selected for each employee: 4 months on the job (i.e., mid-probation), 8 months on the job (i.e., final-probation), and first-annual performance review. This sequential data set allows for a longitudinal perspective of each employee's performance on immediate as well as a long-term criteria which may indicate fluctuations in validity.

Objective Criterion Measure. As mentioned in the Introduction section of this report, multiple sources of criterion data are of immense value in diagnosing idiosyncratic predictor-criterion linkages; linkages which may identify unique relationships which would otherwise have been uninvestigated. Complementary criterion variables account for an incremental portion of the total criterion space, which by definition approximates the entire scope of "job performance." This is contrasted to singular criterion validation efforts which represent an incomplete understanding of an employee's total job performance which can be attributed to criterion deficiency, inclusion of bias, and measurement error.

For the current project, two objective measures of criterion performance were used: tenure and turnover. Tenure will identify the long-term employment status of Eligibility Technicians. This criterion is useful in that it has been determined that the County of San Diego accrues fiscal benefit only after an Eligibility Technician remains in that position for two years. If for any reason the job incumbent leaves the position for a reason other than promotion, turnover results. The criterion of turnover indicates the ability of the selection system to predict employment change which itself has fiscal impact on the County. As such, these objective criteria are more tangible and less biased than supervisory ratings while at the same time demonstrating financial impact on the County in general and the selection system used to hire Eligibility Technicians in specific.

Results

Predictor Analyses

Written Test. This report presents selected results for the purpose at hand; a complete description of all analyses conducted is available (Cesare, Blankenship, Giannetto, & Mandel, 1990). As mentioned previously, the multiple choice test is segmented into six subscales. The descriptive statistics for each subscale and total test score based on applicant ethnicity and gender are presented in Tables 1 and 2 respectively.

In order to test for test score differences based on applicant demographics, a hierarchical multiple regression analysis was conducted using total test score as the dependent variable and applicant ethnicity and gender as the independent variables.

The six ethnic classifications were coded as five dummy-coded vectors with the Filipino classification being coded as zero on all five vectors. These five dummy-coded vectors were entered into the regression equation as a block on the first step. The change in R^2 was used to examine the main effect for the six ethnic groups. All interpretations of significant findings due to applicant ethnicity must be viewed with extreme caution as a result of the extreme disparity in the sample sizes for each group (i.e., 585 blacks vs. 32 American-Indians). Such disparity precludes casual inspection of mean values and subsequent conclusions regarding generalizability. Next, the two gender classifications were coded as one dummy-coded vector with the female condition being coded as zero on that vector. This vector was entered into the regression equation on the second step of the analysis and represents the main effect for applicant gender. This change in R^2 can be interpreted directly as an analysis of covariance (ANCOVA) between the two gender groups partialling out the effects due to applicant ethnicity (Cohen & Cohen, 1983). Finally, the five vectors formed from the cross products of the dummy-coded vectors, for ethnicity and gender, were entered as the final step of the hierarchical regression analysis. Changes in R^2 as a result of the addition of these vectors were used to examine the interaction effects of applicant ethnicity and gender on total test score.

The results of this analyses are shown in Table 3. These results illustrate that applicant ethnicity had a significant effect on total test performance, which accounted for 24% of the total variance. The Scheffe's post-hoc test revealed that: (a) white applicants (M = 81.61) scored significantly different from the Asian (M = 64.73), Filipino (M = 67.35), Hispanic (M = 69.58), and black (M = 71.47) applicants, (b) the mean total test score for American-Indians (M = 76.41) and blacks was significantly different from that obtained by

Table 1

Means of the Subscales and Total Test by Ethnic Group

	Black	Hisp	Asian	Am-Ind	White	Filip
Reading Comprehension	13.45 (2.65)	12.88 (2.93)	11.07 (3.44)	13.97 (2.61)	15.12 (2.43)	11.55 (2.88)
Time Management and Workload Organization	8.63 (1.89)	8.36 (1.88)	7.72 (2.21)	10.03 (1.58)	9.73 (1.85)	7.93 (1.96)
Comprehending and Applying Eligibility Regulations	15.74 (3.57)	15.36 (3.63)	14.75 (3.62)	15.66 (3.40)	17.86 (3.09)	14.65 (3.30)
Computational Skills	6.40 (2.57)	6.59 (2.57)	6.87 (2.35)	7.22 (2.50)	8.00 (1.96)	7.11 (2.38)
Interviewing	13.05 (2.83)	12.56 (3.01)	11.54 (3.55)	14.00 (3.14)	14.90 (2.50)	11.96 (3.10)
Interpersonal Relations	14.20 (3.82)	13.83 (4.03)	12.78 (3.87)	15.53 (3.20)	16.00 (2.81)	14.05 (2.72)
Total	71.47 (11.83)	69.58 (12.79)	64.73 (14.44)	76.41 (12.21)	81.61 (9.94)	67.35 (11.80)

Note. Total N = 3,268. N's for each ethnic group are as follows: black, 585; Hispanic, 591; Asian, 254; American-Indian, 32; white, 1,517; Filipino, 289. Standard deviations for each subscale are presented in parentheses.

Asians and Filipinos, and (c) the mean test score obtained by Hispanics was significantly different from that of Asians.

The results also indicate that applicant gender had a significant effect on test score with males (M = 75.02) outperforming females (M = 74.97). However, the effect due to gender accounted for a negligible proportion of the total variance (i.e., less than 1%) and should be interpreted more in terms of practical

Table 2

Means of the Subscales and Total Test by Gender

	Males	Females
Reading Comprehension	13.86 (3.11)	13.74 (3.01)
Time Management and Workload Organization	8.99 (2.13)	8.96 (2.01)
Comprehending and Applying Eligibility Regulations	16.55 (3.72)	16.45 (3.54)
Computational Skills	7.68 (2.26)	7.13 (2.40)
Interviewing	13.44 (3.14)	13.68 (3.05)
Interpersonal Relations	14.51 (3.73)	15.01 (3.42)
Total	75.02 (13.72)	74.97 (12.83)

Note. Total N = 3,268. N's for each gender are as follows: Males, 953; Females, 2,315. Standard deviations for each subscale are presented in parentheses.

significance rather than statistical significance; this significant finding is more a reflection of the sample size than true differences in test performance based on applicant gender.

An interaction between applicant ethnicity and gender was also found to be significant, though it accounted for less than 1% of the total test score variance. The appropriate cell means are shown in Table 4. Inspection of the cell means focuses attention on (a) the mean score obtained by American-Indian males (M = 83.50) and white males (M = 82.01) against the mean values obtained by all non-white female applicants (blacks, 71.45; Hispanics, 69.27; Asians, 65.62; American-Indians, 72.15; and Filipinos, 65.74), and

Table 3

Hierarchical Multiple Regression Analysis Using Total Test Score as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Applicant Ethnicity	1	.490	.240	206.37**	5,3262			
Applicant Gender	2	.491	.241	172.83**	6,3261	.001	4.16*	1,3261
Ethnicity x Gender	3	.496	.246	96.44**	11,3256	.005	3.87**	5,3256

* $p < .05$. ** $p < .01$.

Table 4

Means of Total Test Score by Applicant Ethnicity and Gender

	Black	Hisp	Asian	Am-Ind	White	Filip
Male	71.54	70.37	63.83	83.50	82.01	70.92
	(12.48)	(13.25)	(14.56)	(10.09)	(10.56)	(10.86)
	145	164	125	12	419	88
Female	71.45	69.27	65.62	72.15	81.44	65.74
	(11.62)	(12.62)	(14.34)	(11.55)	(9.69)	(11.88)
	440	427	129	20	1098	201

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

(b) the difference between the mean score attained by white females (M = 81.44) and the mean score obtained by Asian males (M = 63.83).

Application Blank/Interview Composite. While the application blank and selection interview are administered and evaluated separately, the Department of Social Services aggregates the ratings from each instrument into a composite predictor score which in turn is used to screen applicants from the selection process.

In order to assess the underlying multidimensionality of the application/interview composite, a factor analysis involving a principal components extraction was conducted. Table 5 illustrates the factor loadings

Table 5

Matrix of Orthogonally Rotated Factor Loadings for the Application Blank/Interview Composite

Item	Factor 1	Factor 2
Int5	.779	.222
Int2	.779	.129
Int3	.760	.132
Int1	.747	.189
Int4	.713	.196
App2	.138	.831
App1	.121	.829
App3	.187	.806
App4	.132	.203
Eigenvalue	3.693	1.478
% variance explained	41.0	16.4

Note. N = 1,587.

and summary statistics from the factor analysis of the composite predictor instrument. Two unique factors, identified by instrument method, are evident. Factor 1 is clearly indicative of interview performance; all five interview items loaded above .70 on this factor. The obtained eigenvalue of 3.693 is sufficiently larger than the criterion of 1.00 and in turn accounts for a large proportion of explained variance (i.e., 41%). Similarly, Factor 2 is representative of the application blank instrument with three of the four application blank items loading above .80. Likewise, this factor accounts for a significant proportion (i.e., 16%) of the rating variance. A separate factor score was calculated for each of these two dimensions. Each of these factor scores were used as dependent variables in subsequent analyses.

Table 6 summarizes the R, R^2, R^2 change, and F values for the effects of applicant ethnicity and gender using the Interview factor score as the dependent variable. As can be seen by inspecting the R^2 value and associated F value for the applicant ethnicity variable, the main effect for ethnicity is significant and accounts for almost 2% of the total variance. The cell means for each ethnic classification are shown in Table 7. The Scheffe's post hoc test was unable to detect any mean differences between the ethnic groups. In its place, a Newman-Keuls post hoc test showed that the mean Interview factor score for white applicants (M = .10) was significantly different from that of Hispanic (M = -.16) and Asian applicants (M = -.27). While there was no significant main effect due to applicant gender, there was a significant interaction between ethnicity and gender. The cell means illustrating this interaction are shown in Table 8. It appears that significant differences are present between white males (M = .16) and American-Indian females (M = -.74), and Filipino males (M = -.47), Asian males (M = -.57) and white females (M = .07).

Table 9 summarizes the R, R^2, R^2 change, and F values for the effects of applicant ethnicity and gender using the Application Blank factor score as the dependent variable. As can be seen by inspecting the R^2 value and associated F value for the applicant ethnicity variable, the main effect for ethnicity is significant and accounts for 1% of the total variance. The cell means for each ethnic classification are shown in Table 10. Neither the Scheffe's nor the Newman-Keuls post hoc tests were able to detect any mean differences

Table 6

Hierarchical Multiple Regression Analysis Using the Interview Factor Score as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Applicant Ethnicity	1	.132	.017	4.60**	5,1303			
Applicant Gender	2	.132	.017	3.86**	6,1302	.000	.18	1,1302
Ethnicity x Gender	3	.165	.027	3.32**	11,1297	.010	2.63*	5,1297

* $p < .05$. ** $p < .01$.

Table 7

Means of the Interview Factor Score by Applicant Ethnicity

Black	Hisp	Asian	Am-Ind	White	Filip
-.02	-.16	-.27	-.22	.10	-.20
(1.02)	(.99)	(1.01)	(1.04)	(.99)	(1.09)
134	258	70	11	779	43

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

Table 8

Means of Interview Factor Score by Applicant Ethnicity and Gender

	Black	Hisp	Asian	Am-Ind	White	Filip
Male	.05	-.26	-.57	.07	.16	-.47
	(.95)	(.95)	(.93)	(1.18)	(1.02)	(1.04)
	31	72	33	7	227	20
Female	-.04	-.12	-.00	-.74	.07	.03
	(1.05)	(1.02)	(1.01)	(.50)	(.98)	(1.11)
	103	186	37	4	552	23

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

Table 9

Hierarchical Multiple Regression Analysis Using the Application Blank Factor Score as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Applicant Ethnicity	1	.103	.010	2.80*	5,1303			
Applicant Gender	2	.106	.011	2.31*	6,1302	.001	.81	1,1302
Ethnicity x Gender	3	.124	.015	1.86*	11,1297	.004	1.12	5,1297

* $p < .05$.

Table 10

Means of the Application Blank Factor Score by Applicant Ethnicity

Black	Hisp	Asian	Am-Ind	White	Filip
-.01	-.06	-.12	.07	.09	-.13
(1.03)	(1.07)	(1.03)	(1.09)	(.95)	(.98)
134	258	70	11	779	43

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

between the ethnic groups. Consequently, the less conservative Duncan's post hoc test was then used to compare the set of means obtained on this dependent variable. This test indicated that white applicants (M = .09) performed significantly different from Hispanic applicants (M = -.12). There were no significant effects due to applicant gender or the interaction between ethnicity and gender for this dependent variable.

Criterion Analyses

Mid-Probation Ratings. In an attempt to simplify interpretation of the mid-probation ratings, a factor analysis involving a principal-components extraction and a varimax rotation was conducted to identify the presence of underlying constructs. The results of this analysis are presented in Table 11. This analysis revealed that four factors were extracted from the mid-probation ratings. The first factor appears to be a Work Performance factor, with six items loading above .40 and only one item (number 21) loading highly on a second factor. This factor is characterized by: accuracy, quality, volume, and promptness of performance with minimal supervision and/or under stressful conditions. The second factor appears to be a Flexibility factor with only item 22 loading highly on a second factor. This factor is indicative of cooperation, acceptance of new ideas, performance in new situations, and dealing with the public and/or employees. The

Table 11

Matrix of Orthogonally Rotated Factor Loadings for the Mid-Probation Ratings

Item	Factor 1	Factor 2	Factor 3	Factor 4
11	.704	-.045	.191	.149
12	.675	.152	-.169	.070
18	.674	.362	.109	.126
19	.630	.138	.363	-.085
20	.622	.196	.278	.045
21	.488	.466	.114	-.054
7	.104	.684	.199	-.002
8	.221	.673	.131	.121
13	.020	.607	.025	.182
22	.419	.597	.112	.010
10	.026	.329	.746	-.036
17	.237	-.121	.635	.299
9	.230	.369	.628	.032
14	.094	.025	.164	.808
15	.047	.227	-.031	.808
Eigenvalue	4.666	1.385	1.256	1.127
% variance explained	31.1	9.2	8.4	7.5

Note. N = 491. Items 1 - 6 were not included in this analysis due to limited variability in ratings. Item 16 was not included in this analysis because of its non-relevance to this job class.

third factor is a Workmanship dimension that includes: taking interest in the job, putting forth effort, and doing work neatly. The fourth factor is a Communication dimension, written as well as verbal communication. These four factors correlate .45, .34, .20, and .17 respectively with the overall mid-probation rating; each intercorrelation is significant at the level of $p < .01$.

Final-Probation Ratings. The underlying factor pattern of the final-probation ratings was determined through a factor analysis involving a principal-components extraction and a varimax rotation. The results of this analysis are shown in Table 12 and demonstrate a four factor solution similar to that found for the mid-probation ratings. The first factor is a Work Performance factor, with four items loading above .40. Though not as diverse as the Work Performance factor identified for the mid-probation ratings, this factor is still illustrative of its title; it is representative of: completing a large volume of work promptly, accurately, and neatly. The second factor is defined as an Ability to Work Independently. This factor contains several items which loaded highly on the mid-probation Work Performance factor. In specific, the factor of Ability to Work Independently consists of: performance with minimum supervision, performance under pressure and in new situations, and quality of judgment. The third factor is equivalent to the Flexibility factor found in the mid-probation ratings. This factor is described by: cooperation, acceptance of new ideas, application of effort, and taking interest in one's job. The fourth factor is Communications/Interpersonal Relations which

Table 12

Matrix of Orthogonally Rotated Factor Loadings for the Final-Probation Ratings

Item	Factor 1	Factor 2	Factor 3	Factor 4
19	.729	.261	.136	-.036
11	.699	.192	.024	.135
20	.615	.335	.192	.136
17	.557	-.050	.109	.272
22	.160	.766	.122	-.004
21	.149	.757	.137	-.098
12	.189	.545	.174	.253
18	.453	.467	.142	.131
10	.161	.075	.716	-.001
7	-.042	.172	.711	.127
9	.436	.018	.629	-.051
8	.099	.341	.583	.086
15	.159	.060	.038	.775
14	.440	-.081	-.031	.655
13	-.154	.406	.324	.503
Eigenvalue	4.392	1.589	1.176	1.073
% variance explained	29.3	10.6	7.8	7.2

Note. N = 443. Items 1 - 6 were not included in this analysis due to limited variability in ratings. Item 16 was not included in this analysis because of its non-relevance to this job class.

entails written and verbal communication, and public/employee relations. These four factors correlate .46, .44, .21, and .17 respectively with the overall final-probation rating; each intercorrelation is significant at the level of $p < .01$.

First-Annual Performance Ratings. The underlying factor pattern of the first-annual performance ratings was determined through a factor analysis with a principal-components extraction and a varimax rotation. The results of this analysis are illustrated in Table 13. Despite the loss of one factor, the underlying rating dimensions are quite similar to those produced by the probation ratings. The first factor once again is defined as Work Performance. This factor consists of seven items which load above .40; three of which load highly on the remaining factors. This Work Performance factor is composed of: completing a large volume of work promptly, accurately, and neatly, minimal supervision; it also indicative of a high degree of effort and quality judgment. The second factor is an approximation of the previously defined Flexibility dimension. This factor is composed of six items, which have loadings above .40, only one of which loads highly on another item. For this set of performance ratings, Flexibility is described by: acceptance of new ideas and procedures, performance in new situations and/or when under pressure, cooperation, taking interest in one's job, and public and/or employee relations. The third and final factor once again is Communication, written as well as verbal. The three factors correlate .58, .53, and .16 respectively with the

Table 13

Matrix of Orthogonally Rotated Factor Loadings for the First-Annual Performance Ratings

Item	Factor 1	Factor 2	Factor 3
19	.833	.280	.025
11	.799	.155	.204
20	.751	.255	.123
18	.602	.466	.158
9	.550	.547	-.021
17	.491	.093	.469
12	.473	.388	.354
13	.008	.724	.203
8	.230	.708	.041
22	.360	.654	.273
7	.201	.651	-.097
10	.318	.622	.049
21	.441	.519	.172
15	.084	.121	.845
14	.126	.019	.832
Eigenvalue	6.276	1.599	1.088
% variance explained	41.8	10.7	7.3

Note. N = 251. Items 1 - 6 were not included in this analysis due to limited variability in ratings. Item 16 was not included in this analysis because of its non-relevance to this job class.

overall first-annual performance rating; the first two of which are significant at the level of $p < .01$, while the third is significant at the level of $p < .05$.

The final statistical analysis conducted on the criteria is summarized in Table 14. This analysis shows the degree of intercorrelation between the overall performance ratings for each point in time. As can be seen, there are strong relationships between temporal criteria; relationships which dissipate with time. For example, the intercorrelation between mid-probation ratings and final-probation ratings is .35. In turn, the relationship between mid-probation ratings and first-annual performance is reduced to .16. The same gradient is also evident for final-probation ratings. The very strong correlation between the two annual performance ratings suggests some degree of behavioral consistency in incumbent work performance; although common rater bias and criterion contamination are also legitimate explanations which deserve investigation. Nevertheless, in general it is clear that incumbents receive similar performance ratings over time.

Table 14

Intercorrelation Matrix of Overall Performance Ratings

	Mid Probation	Final Probation	First-Annual Performance
Mid-Probation	1.00	.35 * (482)	.16 * (273)
Final-Probation		1.00	.29 * (276)
First-Annual Performance			1.00

Note. N's appear in parentheses.

* $p < .01$.

Validation Analyses

Performance Rating Criteria. Separate hierarchical multiple regression analyses were conducted, using each of the above mentioned criteria as the dependent variables, to assess the effects of the incumbent demographics and selection instruments.

The results of the multiple regression analysis using the overall mid-probation rating as the criterion are shown in Table 15. As can be seen, neither of the incumbent demographics had a significant effect on this measure of job performance. Similarly, neither of the selection instruments was able to demonstrate a statistically significant effect.

The results of the same analysis using the mid-probation factor of Work Performance as the criterion are summarized in Table 16. Once again, the incumbent demographics failed to produce a significant effect on the dependent variable. While the effects of the demographics and application/interview composite were non-significant, the written test was found to be statistically significant and accounted for an additional 3% of the criterion variance.

Table 17 summarizes the results using the mid-probation factor of Flexibility as the dependent variable. This set of results showed that only one entry had a significant effect on Flexibility: A main effect for incumbent gender which indicated that male incumbents (M = .14) were rated as being significantly more flexible than female incumbents (M = -.14).

Table 15

Hierarchical Multiple Regression Analysis Using the Overall Mid-Probation Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.067	.004	.37	5,408			
Incumbent Gender	2	.105	.011	.76	6,407	.007	2.68	1,407
Ethnicity x Gender	3	.133	.018	.66	11,402	.007	.55	5,402
Test Score	4	.160	.025	.87	12,401	.007	3.15	1,401
Composite	5	.183	.033	1.07	13,400	.008	3.29	1,400

Table 16

Hierarchical Multiple Regression Analysis Using the Mid-Probation Work Performance Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.064	.004	.29	5,364			
Incumbent Gender	2	.080	.006	.39	6,363	.002	.86	1,363
Ethnicity x Gender	3	.108	.012	.38	11,358	.006	.38	5,358
Test Score	4	.202	.041	1.26	12,357	.029	10.84*	1,357
Composite	5	.206	.042	1.21	13,356	.001	.56	1,356

* $p < .01$.

Table 17

Hierarchical Multiple Regression Analysis Using the Mid-Probation Flexibility Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.106	.011	.83	5,364			
Incumbent Gender	2	.160	.025	1.59	6,363	.014	5.34*	1,363
Ethnicity x Gender	3	.195	.038	1.28	11,358	.013	.91	5,358
Test Score	4	.195	.038	1.17	12,357	.000	.01	1,357
Composite	5	.197	.039	1.11	13,356	.001	.39	1,356

* $p < .05$.

The results of the analysis using the mid-probation factor of Workmanship are presented in Table 18. The results illustrate that all five entries into the regression equation produced non-significant effects on the criterion of Workmanship.

Table 18

Hierarchical Multiple Regression Analysis Using the Mid-Probation Workmanship Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.152	.023	1.73	5,364			
Incumbent Gender	2	.159	.025	1.58	6,363	.002	.86	1,363
Ethnicity x Gender	3	.168	.028	.95	11,358	.003	.21	5,358
Test Score	4	.169	.029	.88	12,357	.001	.14	1,357
Composite	5	.170	.029	.82	13,356	.000	.13	1,356

The final regression analysis conducted on the mid-probation criteria focused on the factor of Communication. The results shown in Table 19 indicate that incumbent ethnicity and gender did not have a

significant effect on this dependent variable. However, the interaction between these two variables did produce a significant effect. The cell means for this interaction are presented in Table 20. It appears that black females (M = .24) and American-Indian females (M = .38) are statistically different from Filipino males (M = -.73), Hispanic males (M = -.59), and Asian males (M = -.83). The admonishment of deriving

Table 19

Hierarchical Multiple Regression Analysis Using the Mid-Probation Communication Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.145	.021	1.57	5,364			
Incumbent Gender	2	.171	.029	1.83	6,363	.008	3.09	1,363
Ethnicity x Gender	3	.250	.062	2.17*	11,358	.033	2.53*	5,358
Test Score	4	.251	.063	1.99*	12,357	.001	.09	1,357
Composite	5	.252	.064	1.86*	13,356	.001	.36	1,356

* $p < .05$.

meaningful interpretations from this data, which has been given repeatedly throughout this report, is even more meaningful at this point because the cell sizes are highly disparate (i.e., 1 American-Indian female and 152 white females).

Table 20

Means of the Mid-Probation Communication Factor by Incumbent Ethnicity and Gender

	Black	Hisp	Asian	Am-Ind	White	Filip
Male	-.05	-.59	-.83	.20	.02	-.730
	(.26)	(1.22)	(1.76)	(.39)	(1.12)	(1.25)
	8	23	10	2	67	4
Female	.25	.06	-.16	.38	-.00	-.02
	(1.00)	(.49)	(.27)	(.00)	(.89)	(.09)
	26	79	11	1	152	9

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

The next set of regression analyses were conducted on the five criteria associated with the final-probation ratings. The results of the analyses involving the overall final-probation ratings are summarized in Table 21. For this analysis, each of the three factors associated with the incumbent demographics failed to reach statistical significance. In contrast, the written test was found to be a significant predictor of final-probation performance rating. The predictor composite had a non-significant effect. —

Table 21

Hierarchical Multiple Regression Analysis Using the Overall Final-Probation Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.087	.007	.57	5,372			
Incumbent Gender	2	.088	.008	.49	6,371	.001	.10	1,371
Ethnicity x Gender	3	.166	.027	.94	11,366	.019	1.48	5,366
Test Score	4	.251	.063	2.05*	12,365	.036	13.89**	1,365
Composite	5	.268	.072	2.17*	13,364	.009	3.38	1,364

* $p < .05$. ** $p < .01$.

The results for the final-probation factor of Work Performance are presented in Table 22. The demographic variables had a negligible effect on this dependent variable. In terms of the selection instruments, the test had a significant effect while the application/interview predictor did not.

Table 22

Hierarchical Multiple Regression Analysis Using the Final-Probation Work Performance Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.160	.026	1.77	5,337			
Incumbent Gender	2	.160	.026	1.48	6,336	.000	.01	1,336
Ethnicity x Gender	3	.217	.047	1.49	11,331	.021	1.48	5,331
Test Score	4	.242	.058	1.71	12,330	.011	3.99*	1,330
Composite	5	.242	.058	1.58	13,329	.000	.81	1,329

* $p < .05$.

As can be seen in Table 23, each of the five entries into the regression equation failed to reach statistical significance in predicting the dependent variable of Ability to Work Independently.

Table 23

Hierarchical Multiple Regression Analysis Using the Final-Probation Ability to Work Independently Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.107	.011	.78	5,337			
Incumbent Gender	2	.107	.011	.65	6,336	.000	.02	1,336
Ethnicity x Gender	3	.177	.031	.98	11,331	.020	1.36	5,331
Test Score	4	.184	.034	.96	12,330	.003	.77	1,330
Composite	5	.185	.034	.89	13,329	.000	.17	1,329

The results of the analysis using the factor of Flexibility as the dependent variable are shown in Table 24. As evidenced by the ethnicity main effect, different ethnic groups were rated differently on this measure. The cell means for each of the ethnic groups are presented in Table 25. Neither the Scheffe's, the Newman-Keuls, nor the Duncan's post hoc test could detect any significant differences between the group means. However, it is believed that the mean for the white incumbents (M = .09) was significantly different from that of the Filipino incumbents (M = -.66).

Table 24

Hierarchical Multiple Regression Analysis Using the Final-Probation Flexibility Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.181	.033	2.27*	5,337			
Incumbent Gender	2	.195	.038	2.20*	6,336	.005	1.81	1,336
Ethnicity x Gender	3	.253	.064	2.06*	11,331	.026	1.86	5,331
Test Score	4	.262	.068	2.03*	12,330	.004	1.63	1,330
Composite	5	.266	.071	1.93*	13,329	.003	.82	1,329

* $p < .05$.

Table 25

Means of the Final-Probation Flexibility Factor by Incumbent Ethnicity

Black	Hisp	Asian	Am-Ind	White	Filip
-.17	.01	-.27	-.36	.09	-.66
(1.03)	(1.04)	(.80)	(1.13)	(.98)	(.68)
29	92	22	4	207	10

Note. Standard Deviations appear in parentheses. Number of applicants per cell is listed beneath the Standard Deviations.

The last set of regression analyses conducted on the final-probation ratings used the factor of Communications/Interpersonal Relations as the dependent variable. The results of this analysis are summarized in Table 26 which illustrates that incumbent gender was the only variable which had a significant effect in predicting this dependent variable. The main effect for gender was based on the fact that female incumbents received a higher mean (M = .09) value on Communications/Interpersonal Relations than did male incumbents (M = -.19). As is shown, neither the test nor the application/interview composite had a significant effect on this criterion.

Table 26

Hierarchical Multiple Regression Analysis Using the Final-Probation Communication/Interpersonal Relations Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.092	.008	.57	5,337			
Incumbent Gender	2	.168	.028	1.64	6,336	.020	6.88*	1,336
Ethnicity x Gender	3	.184	.034	1.06	11,331	.006	.38	5,331
Test Score	4	.185	.034	.97	12,330	.000	.05	1,330
Composite	5	.200	.040	1.05	13,329	.006	2.01	1,329

* $p < .01$.

The results of the multiple regression analysis using the overall first-annual performance rating as the dependent variable are presented in Table 27. Each of the five predictor variables failed to reach a degree of statistical significance in predicting first annual ratings of job performance.

Table 27

Hierarchical Multiple Regression Analysis Using the Overall First-Annual Performance Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.163	.027	1.22	5,223			
Incumbent Gender	2	.180	.033	1.24	6,222	.006	1.35	1,222
Ethnicity x Gender	3	.211	.044	.91	11,217	.011	.54	5,217
Test Score	4	.220	.048	.92	12,216	.004	.94	1,216
Composite	5	.229	.052	.92	13,215	.004	.92	1,215

Table 28 shows the results of the analysis which were conducted using the first-annual performance factor of Work Performance as the dependent variable. In this analysis, the written test was found to have a significant effect on the criterion. Despite this degree of statistical significance, the written test accounted for only 2% of the rating variance. The application/interview composite was once again non-significant.

Table 28

Hierarchical Multiple Regression Analysis Using the First-Annual Performance Rating Work Performance Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.183	.034	1.39	5,200			
Incumbent Gender	2	.186	.034	1.18	6,199	.000	.16	1,199
Ethnicity x Gender	3	.226	.051	.95	11,194	.017	.68	5,194
Test Score	4	.266	.071	1.22	12,193	.020	4.03*	1,193
Composite	5	.266	.071	1.13	13,192	.000	.07	1,192

* $p < .05$.

The inability of the demographic variables or the selection instruments to predict the dependent variable of Flexibility are summarized in Table 29. As can be seen, neither the applicant demographics, the written test, nor the application/interview composite had a significant effect on this dependent variable. The written test accounted for approximately 1% of this variance, while the application/interview composite had no additive effect on the proportion of variance accounted for.

Table 29

Hierarchical Multiple Regression Analysis Using the First-Annual Performance Rating Flexibility Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.087	.007	.31	5,200			
Incumbent Gender	2	.088	.007	.25	6,199	.000	.00	1,199
Ethnicity x Gender	3	.235	.055	1.03	11,194	.048	1.95	5,194
Test Score	4	.254	.064	1.11	12,193	.009	1.94	1,193
Composite	5	.254	.064	1.02	13,192	.000	.19	1,192

A similar pattern of non-significant effects are presented in Table 30 which show that none of the variables were able to predict the criterion of Communication at the point in time of the first-annual performance review.

Table 30

Hierarchical Multiple Regression Analysis Using the First-Annual Performance Rating Communication Factor as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Incumbent Ethnicity	1	.215	.046	1.94	5,200			
Incumbent Gender	2	.215	.046	1.61	6,199	.000	.00	1,199
Ethnicity x Gender	3	.303	.092	1.78	11,194	.046	1.95	5,194
Test Score	4	.308	.095	1.69	12,193	.003	.68	1,193
Composite	5	.308	.095	1.55	13,192	.000	.02	1,192

Test Fairness. According to the *Uniform Guidelines on Employee Selection Procedures* (1978), the definition of test fairness is that advocated by Cleary (1968). This model of selection fairness employs an empirical comparison of the regression slopes and intercepts for the various subgroups (Cascio, 1987).

In terms of the overall mid-probation performance rating criterion, there were no significant differences between the regression slopes or intercepts for two incumbent demographic variables: ethnicity ($F(8,438) = .52, p > .05$) and minority status ($F(2,444) = .61, p > .05$). Thus, a common regression is appropriate in predicting job performance for both aspects of determining an incumbent's ethnic status. However, a significant difference was found for gender ($F(2,444) = 3.75, p < .05$). Post-hoc analysis showed that the significant degree of predictive unfairness was directed at males ($F(1,132) = 7.97, p < .01$) and not females ($F(1,312) = .24, p > .05$), as is commonly alleged.

This element of predictive discrimination against males disappears when the criterion of overall final-probation performance rating is used as the criterion ($F(2,432) = .78, p > .05$). In similar form, there are common regression lines and no slope discrepancies when comparing the six ethnic groups ($F(8,426) = .76, p > .05$) or the two minority status partitions ($F(2,432) = .61, p > .05$). In terms of final-probation ratings, there is no aspect of test unfairness for any incumbent subpopulation.

Similar findings are also evident for the criterion of overall first-annual performance rating. Neither the two conditions of minority status ($F(2,240) = .96, p > .05$) nor incumbent gender ($F(2,226) = 1.10, p > .05$) had a significantly different slope or intercept. Consequently, there was no degree of test unfairness present at this criterion point.

Corrected Validation Coefficients. This section of the report summarizes the psychometric corrections that were made to the observed validity coefficients described earlier. Two basic corrections were applied to the observed validity coefficients. The first was a correction for the effects of range restriction due to explicit and incidental selection on the two selection instruments (i.e., written test and application/interview composite). The second correction was made to adjust for unreliability in the criteria (i.e., overall performance ratings at each of the four criterion points in time).

It is important to remember that the selection process used to hire Eligibility Technicians is conducted in two steps. All candidates who meet the minimum qualifications for the job class are required to take and pass the written multiple-choice test. Candidates who pass this exam are then placed on an eligibility list in terms of their rank ordered performance on the test. As candidates are certified to the Department of Social Services, they are given ratings based on their application blank and selection interview; these ratings are averaged and combined to form the predictor composite.

Table 31 shows the intercorrelations between the predictors and overall mid-probation rating. After correcting for restriction in range, the correlation between the predictor composite and test score increase from .18 to .33; this coefficient was not corrected for criterion unreliability. The correction for restriction in range also improved the correlations between test score and performance rating from .10 to .20, and predictor composite and performance rating .09 to .11. Once this correction was made, these two correlations were then corrected for criterion unreliability. This correction increased the correlation between test score and performance rating from .20 to .26 and the correlation between predictor composite and criterion rating from .11 to .14.

These corrected coefficients were then re-entered into a hierarchical multiple regression equation similar to that described previously. Total test score was entered on the first step, followed by the application/interview composite on the second step. The results of this regression analysis are summarized in Table 32. Once again, the results demonstrate that the written test had a significant predictive effect on the overall mid-probation ratings. In addition to the statistical significance, the written test accounted for nearly 7% of the total rating variance. Conversely, the predictor composite failed to produce a significant effect on the criterion and accounted for a negligible proportion of variance.

Table 31

Uncorrected Intercorrelation Matrix of Selection Methods and Overall Mid-Probation Ratings

	(1)	(2)	(3)
(1) Test Score	1.00	.18** (1338)	.10* (449)
(2) Predictor Composite		1.00	.09* (485)
(3) Performance Rating			1.00

Note. Values shown in parentheses represent the sample size.

* $p < .05$. ** $p < .01$.

Table 32

Hierarchical Multiple Regression Analysis Using the Corrected Coefficients and Overall Mid-Probation Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Test Score	1	.260	.067	32.41*	1,447			
Composite	2	.266	.071	17.01*	2,446	.004	1.58	1,446

* $p < .01$.

Table 33 is a summary of the intercorrelations between the two predictor instruments and overall final-probation ratings. Identical to the results found in Table 31, the correction for restriction in range raised the correlation between test score and predictor composite from .18 to .33. This correction raised the correlation between test score and final-probation rating from .14 to .26. Likewise, the correlation between the predictor composite and the final-probation rating was increased from .07 to .13. Following these corrections, the criterion variable of final-probation ratings was then corrected for unreliability. This correction improved the correlation between test score and criterion from .26 to .34, and between predictor composite and criterion from .13 to .17.

Table 33

Uncorrected Intercorrelation Matrix of Selection Methods and Overall Final-Probation Ratings

	(1)	(2)	(3)
(1) Test Score	1.00	.18* (1338)	.14* (409)
(2) Predictor Composite		1.00	.07 (429)
(3) Performance Rating			1.00

Note. Values shown in parentheses represent the sample size.

* $p < .01$.

Similar to the procedure just described for the mid-probation ratings, the corrected coefficients were entered into a hierarchical multiple regression analysis, the results of which are summarized in Table 34. The written test showed a significant relationship with the final-overall probation rating. Employees who performed well on the test also performed well on the job. This relationship accounted for nearly 12% of the variance. The application/interview composite was found to be non-significant; it had no predictive value concerning final-probation ratings.

Table 34

Hierarchical Multiple Regression Analysis Using the Corrected Coefficients and Overall Final-Probation Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Test Score	1	.340	.115	53.20*	1,407			
Composite	2	.345	.119	27.51*	2,406	.004	1.73	1,406

* $p < .01$.

Table 35 represents the intercorrelations between the written test, the application/interview composite, and first-annual performance ratings. Correcting for restriction of range increase the correlation between test score and first-annual performance rating from .07 to .13. Likewise, the correlation between the predictor composite and first-annual performance ratings went from -.05 to -.02. Subsequently, these coefficients were corrected for unreliability in the criterion rating. This correction improved the correlation between the test score and the criterion from .13 to .17. The impact of this correction had little effect on the relationship between the predictor composite and criterion, changing it from -.02 to -.03.

Table 35

Uncorrected Intercorrelation Matrix of Selection Methods and Overall First-Annual Performance Ratings

	(1)	(2)	(3)
(1) Test Score	1.00	.18* (1338)	.07 (250)
(2) Predictor Composite		1.00	-.05 (237)
(3) Performance Rating			1.00

Note. Values shown in parentheses represent the sample size.

* $p < .01$.

The results of the hierarchical multiple regression analysis using the corrected correlation coefficients and first-annual performance ratings are shown in Table 36. As indicated, the written test was found to be a significant predictor of job performance at the time of first-annual performance review, demonstrating a positive relationship. Though it was a statistically significant effect, it accounted for less than 3% of the total criterion variance. The application/interview composite was found to be non-significant and accounted for a negligible amount of variance.

Criterion Composite. Composite ratings were calculated by summing the individual ratings given on the 15 relevant performance evaluation dimensions (see Table 11), for each of the three criterion time frames (e.g., mid-probation, final-probation, and first-annual performance review). The composite was then regressed onto each applicant's total test score. These composite ratings were corrected for restriction in range and criterion unreliability and then correlated with the written test.

For the mid-probation time period, a significant relationship between total test score and the criterion composite ($F(1,438) = 57.10, p < .01$) was evident. The R for this relationship was .34 and accounted for almost 12% of the total variance. Thus, the written test has a significant degree of relationship with the multidimensional performance composite measured midway through an Eligibility Technician's probationary training.

A similar pattern of results were also evident for the final-probation and first-annual performance

Table 36

Hierarchical Multiple Regression Analysis Using the Corrected Coefficients and First-Annual Performance Rating as the Dependent Variable

Variables Entered	Step Number	Multiple R	Cumulative R^2	F	df	R^2 Change	F	df
Test Score	1	.170	.029	6.99**	1,235			
Composite	2	.193	.037	4.52*	2,234	.008	2.02	1,234

* $p < .05$. ** $p < .01$.

review criteria. A significant relationship was also found between total test score and the final-probation composite ($F(1,402) = 31.61, p < .01$), with an R of .27, which accounted for approximately 7% of the total variance. Likewise, total test score was also able to predict the composite of first-annual performance ratings ($F(1,235) = 56.42, p < .01$). The R for this relationship was .44 and accounted for approximately 20% of the total rating variance. Clearly, the written test shows a strong pattern of relationship with this criterion, when evaluated at three different time intervals over a one-year period. Once again, the written test has a longitudinal ability to measure job performance significantly.

Tenure. Length of service, as an Eligibility Technician for the County of San Diego, was computed for each employee who left the job class. This criterion was then correlated with the written test score to give an understanding of the test's ability to predict tenure. The uncorrected Pearson correlation coefficient was found to be -.16. After correcting for restriction of range in the predictor, the correlation became -.31 ($p < .01$). These results indicate an inverse relationship between test score performance and tenure: Employees who scored well on the test left the job class of Eligibility Technician sooner than those employee who scored lower on the written test.

Turnover. For the purposes of this report, turnover was defined in dichotomous terms: satisfactory and unsatisfactory. Satisfactory turnover includes resignation, retirement, and expiration of appointment. Conversely, unsatisfactory termination represents instances such as position abandonment, dismissal during probation or employment, and failure to return from leave or suspension.

The point-biserial correlation between an employee's score on the written test and his/her termination/resignation classification was found to be .09. This value increased to .17 after correcting for restriction of range due to selection on the written test ($p < .01$). In explanation, employees who score poorly on the written test tend to leave their position as an Eligibility Technician for unsatisfactory reasons while employees who score higher on the test are likely to terminate their appointment for more satisfactory causes.

A discriminant analysis was also conducted to investigate the ability of the written test to classify termination/resignation correctly. The results indicated that the written test was successful in classifying 63% of unsatisfactory turnover and 59% of those employees who left the position based on satisfactory reasons. In general, the written test correctly classified 59% of employees into their proper groups.

Discussion

Overview

The written test and application blank/interview predictor composite were found to possess appropriate degrees of internal reliability. A factor analysis with Principal-Components extraction and varimax rotation found consistent factor structures for each of the three sets of supervisory ratings. Across a majority of the Work Performance factor scores, the written test was shown to be a successful predictor of subsequent job performance; individuals who performed well on the test, in turn performed well on the job. The predictor composite failed to predict any of the criteria. The written test was also able to demonstrate a significant relationship with two objective indices: tenure and turnover. And finally, the predictors were found to be fair and thus did not discriminate against any protected class.

Predictor Issues

Written Test. The predictive validity of this instrument cannot be denied. Total test score was able to predict the singular most important criterion, the factor score of Work Performance, at the time of mid-probation, final-probation, and first-annual performance review. A significant and positive relationship was evident: Individuals who scored well on the test performed better on the job than those who scored poorly. Furthermore, the test did not discriminate against any ethnic class or gender protected by the Civil Rights Act of 1964.

Noticeable differences are present at the level of the individual subscales. First, the subscale of Time Management and Workload Organization lacks sufficient reliability to make it useful. Relative to the other subscales, Time Management and Workload Organization is the most difficult; on the average, only 63% of items are answered correctly. Further analysis should be conducted to examine if the items are worded ambiguously, written at a proper level of reading ability, and if there is any systematic bias inherent within the distractors. Beyond this point, the lack of reliability may also be attributed to improper content sampling, homogeneity of the domain being sampled, test length, and/or order effects (Anastasi, 1982).

Improper content sampling suggests that a subscale may be measuring only a small portion of the entire domain or that it possesses a disproportionate number of items on a topic relative to its importance. Similarly, domain homogeneity represents the degree to which the items are in fact measuring the same content area (i.e., dimension) or if multiple dimensions are present. Clearly, a homogeneous domain produces higher reliability than one which is heterogeneous. Whether it is improper sampling or misplaced items, each problem will produce error variance; variance which minimizes internal consistency. This measurement error underscores the lack of predictive validity in that an empirically-derived validity coefficient cannot exceed the square root of its corresponding reliability.

Situations such as these must be investigated by a panel of trained subject matter experts in order to determine the specific type of problem and resolution. Typically, this involves the process of "retranslation"; subject matter experts individually classify test items into the content dimensions which they believe they measure (Smith & Kendall, 1963). Once each of the experts has done this, the panel reconvenes to discuss their judgments, inconsistencies, and clarify the definition of each dimension. This process determines the inter-rater reliability for each subscale; a measure which may complement a post-hoc determination of subscale reliability.

Though untested empirically, yet worthy of investigation, is the notion that the poor reliability of the Time Management and Workload Organization subscale is symptomatic of its ordinal position within the entire test. Perhaps some systematic variation occurs during the entire test process. This variation may also be affecting other subscales as well (i.e., contrast effect, spurious score inflation), albeit without negative effects to date. A redistribution of the subscale sequence would be appropriate, insightful, and practical. A point such as this is illustrative of the type of monitoring and revision that must be done to ensure fairness and maximize the psychometric quality of the total test and its individual components.

The final point regarding the written test in general and the subscales in particular is the aspect of cost effectiveness or marginal utility. While a simultaneous research project has already demonstrated cost savings of $6.9 million to the County (Blankenship & Cesare, 1990), an identical procedure should be applied to each of the subscales in order to determine their relative fiscal standing. Though the present report summarizes the psychometric properties of the subscale, a scale-based utility analysis would crystallize their predictive strengths and weaknesses on a dollar metric; a metric which is commonly understood by practitioners as well as researchers. Such an assessment would provide a sound and quantifiable standard by which subsequent interventions could be evaluated.

Application/Interview Composite. Just as the predictive ability of the written test is undeniable, the results for the application/interview composite are equally definitive: This instrument has no predictive effect. Despite the fact that the factor structure of the instrument is defined completely by the two types of items (i.e., interview vs. application) and that the respective instruments demonstrate adequate reliability, no ability to predict job performance is manifest. Though uninvestigated, it's marginal utility is certain to be negligible. The individual components of the composite will be addressed successively.

Application Blank. As stated in the introduction section of this report, the empirical literature on training and experience applications has shown that the average uncorrected criterion-related validity coefficient is .10; .13 after correcting for unreliability in the criterion (Hunter & Hunter, 1984). Moreover, a recent meta analysis revealed that the magnitude of a predictive validity coefficient may be the product of the scoring method used (McDaniel, Schmidt, & Hunter 1988). Those authors argued forcefully for additional investigation into the exact content being measured by a training and evaluation application. Several corrective approaches for the data at hand follow.

The first suggestion centers on the availability of a similar, albeit more valid, training and experience application from an equivalent like-sized County or local government. Given the prevalence of Eligibility Technicians, or synonymous job classes, this option is highly probable. Obviously, this acquisition minimizes cost in terms of time, labor, and research. While issues relevant to situational specificity, criterion linkages, and pilot testing would have to be addressed, the likelihood of major modification is low.

A second alternative is to revise the instrument which is currently in use. This would necessitate the involvement of subject matter experts to identify the anticipated criterion space, design appropriate job-related items, and an accurate scoring system. It is important to understand the conceptual link between predictors and criteria within a selection system: Ideally each predictor would account for additional unique variance in the criterion space; variance overlap minimizes predictive accuracy. Specific to the present selection system at hand, the written test consistently demonstrated the ability to predict the derived factor score of Work Performance while it was equally unable to predict the criteria of Flexibility or the Ability to Work Independently; the revised composite should be structured to predict these latter criteria. The same point can be made for single items, or subsets of items, within the selection instrument. A selection system premised on hypothesized relationships between predictors and criteria clarifies their respective roles, enhances prediction quality, and allows for minor modification periodically, rather than complete revision similar to that suggested here.

As was just mentioned, another solution is to revise the existing instrument. This requires significant planning, research, and pilot testing on the part of the County. It is highly likely that the current application lacks the necessary quantity of items needed to measure an appropriate range of performance. The new application blank should be lengthier; possessing items that assess additional components of the applicant's training and experience. Instead of just focusing on global indices, the revised instrument should examine a variety of subtopics relevant to an applicant's work/education history. For example, instead of asking if the applicant has had supervisory experience, a series of related questions should be asked about that experience: whom did you supervise, how many subordinates, their job titles, what supervisory functions did you engage in (e.g., feedback, discipline, training, mentoring), and a situational question intended to validate the applicant's experience behaviorally in a hypothetical paradigm. This process should be replicated for each

subset of items (i.e., dimension) all the while understanding that this dimension is intended to measure a specific criterion/criteria.

Likewise, the current rating scale used by the raters seems to represent a condensed scope of applicant response. The rating factors must reflect the whole range of possible responses, accurately. This ensures the ability to discriminate between subtle differences in performance; differences which may improve predictive accuracy. It necessarily follows that training be conducted to formalize the correct rating process, foster a consistent understanding of the dimensions, and to receive feedback from the raters themselves (i.e., instrument deficiency). In their review of formalized scoring methods, McDaniel et al. (1988) found the Illinois job element, behavioral consistency, point, and task methods attained respective validity coefficients of .20, .45, .11, and .15. Further investigation into this area by the County is encouraged.

An extension of the training and experience application is the weighted application blank (England, 1971). This selection method serves to distinguish between successful and unsuccessful employees based on a series of items which are weighted to reflect their degree of importance in predicting job success. While this procedure has demonstrated psychometric effectiveness (Cascio, 1987) it has also been demonstrated to have fiscal benefit as well: Lee and Booth (1974) found that a weighted application used in selecting clerical employees saved a county government over $250,000 over two years.

England (1971) has noted that a weighted application blank is particularly effective in those employment situations in which the specified job contains the following characteristics: (a) a large number of employees are performing similar activities; (b) adequate personnel records are available on employees; (c) require long and costly training programs (i.e., probation); (d) high turnover; and (e) large number of applicants relative to position openings. These criteria are consistent with those present for the position of Eligibility Technician and thus warrant additional application to the County of San Diego.

The last topic to be mentioned in relation to application blank-type predictive instruments is the biographical information blank (i.e., biodata). Like the weighted application blank, biodata is a self-report measure, yet differences between the two techniques do exist. Typically, biodata is done through a multiple-choice format, involves a larger volume of items, and in turn includes items that are not covered in the weighted application blank. Frequent reviews of the empirical literature on employee selection have consistently shown biodata to be an effective predictor of various criteria (Hunter & Hunter, 1984; Muchinsky, 1986; Reilly & Chao, 1982). Moreover, work done by Owens and his colleagues (Brush & Owens, 1979; Owens, 1976; Owens & Schoenfeldt, 1979; Schoenfeldt, 1974) has produced a longitudinal taxonomy relating biodata to job performance and satisfaction. Thus, variations of the application blank are able to demonstrate adequate prediction, cost savings, and long-term benefit to a human resource system.

Interview. As was stated in the introduction section of this report, traditional narrative reviews of the literature consistently concluded the interview to be of little use in selecting employees (Arvey & Campion, 1982; Schmitt, 1976). However, a recent meta-analysis found that interviews which were highly structured attained a corrected validity coefficient of .41 (McDaniel et al., 1987). Though highly structured, the interview portion of the predictor composite did not approach that degree of prediction. Perhaps the extreme degree of structure inherent within the interview has a curvilinear relationship with validity: too much or too little is deleterious; a moderate amount may be optimal. This could be assessed using several rating dimensions and quantities of structure. While reliability may increase proportionately with amount of structure (i.e., more objective) validity may be mitigated due to restricted range. Additional investigation on this area should be conducted to clarify this hypothesized relationship and to gain insight into the dysfunctional validity coefficient of the interview itself.

While much of the advice intended to improve the effectiveness of the interview is similar to that outlined above for the application blank it will be reiterated here in brief. It is vital to align the content (i.e., rating dimensions) of the interview with a corresponding segment of the criterion, increase the scope of the subject matter examined within the interview to assess applicant performance more completely, and widen the interview rating scale to allow for maximum variability across performance levels. Interviewer training

must also be made mandatory for all panel members. It is highly likely that improperly trained interviewers may minimize or even nullify the effectiveness of a methodologically sound instrument. Obviously, the degree of interview structure precludes training on topics of content (e.g., illegal questions, non-job relatedness of inquiry). However, training should be given on topics of process (e.g., rating errors, bias, inaccuracy).

This process training would increase interviewer awareness concerning several shortcomings which their ratings appear to substantiate. Incorporation of learning principles such as group discussion, practice-and-feedback, and modeling would likely remedy much of the halo error, central tendency error, and inaccuracy that characterize the present set of interview ratings. Though reduction in error does not imply increased accuracy (Bernardin & Pence, 1980), rating quality will be improved. This element of quality is seen as a major premise for heightened validity.

A popular strategy for enhancing validity is to develop the interview around situational types of questions (Weekley & Gier, 1987). Latham and his colleagues (Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980) have shown the validity of the situational interview to range from .14 to .46. While the existing interview is composed of situational-type questions, neither the actual development of the questions nor the rating scale were developed with the same degree of methodological rigor as that specified by Latham et al. (1980). Typically, this sequence involves the compilation of critical incidents for the focal position; development of questions premised on those incidents; construction of a behavioral rating scale for each question; conducting a pilot test; and performing an item analysis to identify which items have suitable reliability and variance. This process should be done thoroughly for each rating dimension measured by the interview.

Another issue relevant to the selection of Eligibility Technicians also exists. In some cases, interviewers were able to review a candidate's application blank prior to the interview. It is not difficult to imagine that this may produce more bias than accuracy in an interview rating. Perhaps knowledge of a candidate's performance of the application produces an expectancy effect; an effect which serves to contaminate the interviewer's judgment on the subsequent interview (Wyer & Srull, 1981). This effect postulates a spuriously high direct correlation between the scores given on two instruments due to the interviewer's knowledge of the candidates earlier application rating. Clearly, a more appropriate solution is to have the same balanced panels to either review applications or conduct interviews (Dreher, Ash, & Hancock, 1988). The same raters should not evaluate both instruments for the same candidate.

A final point related to the predictor instruments concerns the opportunity for adding an honesty test to the revised selection system. A large aspect of an Eligibility Technician's job is the determination and distribution of financial payment to welfare applicants. It was recently publicized that at least five Eligibility Technicians were suspected of stealing funds from the County (Flynn, 1990). The honesty test is one possible avenue to prevent such misbehavior from reoccurring.

Sackett and Harris (1984) reviewed the paper and pencil versions of several honesty tests and found high reliability, no gender/race differences, and in general positive correlations. By the same token, continued study in this area was encouraged. While adjoining issues need to be addressed to clarify the role of honesty tests, suitable predictors must be installed into the system to measure and predict that appropriate portion of the criterion space. If this selection void remains, naive prediction will continue and the probability of unethical behavior will be evident; behavior which is costly to the County immediately and over time (e.g., poor publicity, misdirected aid, federal sanctions).

Criterion Issues

Perhaps the most critical step in the entire process of employee selection is the development of a relevant, reliable, accurate, and comprehensive criterion. As stated earlier, if measurement of the criterion is poor (e.g., contaminated, deficient, distorted, or vulnerable to scale unit bias) then subsequent selection decisions will also be poor (Brogden & Taylor, 1950).

An initial reaction to problems associated with performance ratings is to consider an alternative performance appraisal format. Indeed, a job-specific instrument for Eligibility Technicians would be an improvement over the standard graphic rating scale used by the County. The development of a new instrument should be behaviorally based, in correspondence with the job analysis, dimension specific, and accepted by the raters (Banks & Murphy, 1985). However, research has shown that no particular format is superior to the others (Kingstrom & Bass, 1981). Likewise, it is vital to remember that a rating format alone is not capable of producing a valid or accurate rating. The point to be made is this: A new performance measurement format is a necessary, not a sufficient quality for a professional performance measurement system.

Just as inconclusive as the research on rating formats has been, the research on rater training has been equally definitive: Rater training does improve the psychometric quality and accuracy of performance measurement (Smith, 1986). A series of rater training workshops should be provided for each supervisor, whose attendance must be made mandatory. One typical facet of a supervisors job description is providing feedback to subordinates; thus, it is a stated function of their job and as such their participation in training is necessary. The series of workshops could address: dimension definitions, appropriate use of performance standards as listed by the anchors on the rating scale, differentiation between relevant vs. irrelevant performance, common rating errors, rating accuracy, behavioral documentation, and providing effective feedback to subordinates. Although this system of training seems comprehensive and implicitly costly, it is wise to remember that expense such as this is designed not only to improve the quality of job performance measurement, but also of job performance itself. While an inaccurate performance appraisal produces many detrimental outcomes, an accurate performance rating fed back to a subordinate correctly serves to reinforce behavioral strengths and identify areas of weakness which are to be alleviated through training, effort, or coaching, eventuating in improved individual and organizational performance.

Another suggestion designed to improve the quality of the performance appraisal criterion is to modify the existing system to allow for multiple raters to contribute in the process. Multiple raters would produce more variance in the criterion, minimize the element of rater bias, evoke greater acceptance by ratees, and increase the quality and quantity of feedback received by the subordinate. This method of assessment also serves as a foundation for the adoption of a multitrait-multirater (Lawler, 1967) paradigm of construct validity which will be addressed in the next section of this report. Research has shown that self ratings tend to be more lenient than supervisory ratings (Mabe & West, 1982); yet also possess less evidence of halo error than supervisory ratings (Heneman, 1974); while peer ratings are typically of high quality and accuracy (Imada, 1982; Kane & Lawler, 1978).

Recently, Harris and Schaubroeck (1988) conducted a meta analysis of the interrelationship between self, peer, and supervisory ratings. Their results indicated a high correlation between peer and supervisory ratings (.62), but only moderate correlations between self and supervisor ratings (.35) and self and peer ratings (.36). This states convincingly that differential perspectives of ratee performance do exist. Although these correlations are relatively low, they should not be dismissed as useless. Different perspectives such as these are insightful into performance feedback (Ilgen, Fisher, & Taylor, 1979), possible job redesign (Hackman & Oldham, 1976), and goal setting (Locke, Shaw, Saari, & Latham, 1981). Moreover, it is intuitively important to recommend that divergent perspectives demand additional research than those which are highly congruent.

The effectiveness of the objective criterion indices was also a partial strength of this study. It was originally intended to use individualized error rates as the objective criterion measure. Unfortunately, these data were not made available and as such remain unanalyzed. Despite this omission, the written test still demonstrated significant relationships with tenure and turnover. Of paramount importance is the directionality of the relationships between these variables: An inverse relationship between test score and tenure and a positive relationship between the test and turnover.

The test-turnover relationship warrants further investigation. Clearly, high performers on the test

are leaving the job class quicker than those who score less well on the test. The intrinsic qualities of the job class deserve additional examination. For example, the stressfully high work load coupled with low pay serves as incentive to leave the job class as soon as possible. Job enrichment or redesign, merit-based compensation, a redefined career ladder, and enhanced skills training represent only a few of many recommendations which may minimize the short tenure associated with high test performers.

Validation Issues

This project was designed as a criterion-related validation study; utilizing a predictive design which includes use of a selection instrument. The results of this report indicate an uncorrected validity coefficient of .18. This value compares favorably with that found in a recent meta-analysis of validation studies which found similar studies to have an average validity coefficient of .26 (Schmitt et al., 1984). It should also be noted that the mean validity coefficient of .26 is equivalent to those values found for predictive designs without selection (.29) and concurrent designs (.34). The validity coefficient of .18 found in this report should also be compared against the mean validity coefficient of those studies which also used performance ratings as the criterion measure (.26).

Clearly, a strength of this study was the large sample size employed for the mid-probation (449), final-probation (409), and first-annual (250) criterion time periods. Numbers of this size neglect the problems associated with smaller more typical validation samples (Lent et al., 1971; Schmidt & Hunter, 1980). Likewise, Schmidt, Hunter, and Urry (1976) demonstrated that, under typical parameters of most validation studies, a valid relationship will only be detected in 50% of the cases. In contrast, the current study possesses a suitable sample size and its cognate statistical power.

On the other hand, the often mentioned weakness of this study was the unequal distribution of ethnic class members. While this shortcoming is not dismissed, precautions were taken. First, data were gathered on every applicant and all subsequent employees through the years of 1986 and 1990. This method of sampling represents an optimal solution to an otherwise tenuous situation. Exhaustive data collection precludes doubt of sampling error, misrepresentation, and biased results; rather it represents factual evidence instead of postulated inference. No element of accuracy was withheld.

One major, albeit implicit, problem in conducting a predictive validation study is the element of time required to gather an appropriate sample on both the predictor and criterion variables. It is recommended that immediate content validation studies be conducted until suitable data aggregation necessary for a criterion-related study has occurred. This recommendation is even more compelling given the very real possibility of proposed legislation aimed at overturning recent Supreme Court decisions (e.g., Wards Cove Packing v. Atonio, 1989) which lessened the employer's burden of responsibility (see Kleiman & Faley, 1978 for a review of legal cases concerning content validity). While typically conceptualized as a purely subjective index, several quantifiable methods do exist. Two of these methods will be presented.

Content Validity. While a review of content validation is well beyond the scope of this report, the five-step framework developed by Mussio and Smith (1973) is presented as a general outline: (1) conduct a thorough and reliable job analysis, (2) construct the examination, (3) calculate the job relevance of the examination, (4) establish the cutoff scores for the content valid examination, and (5) select the raters for conducting the content validity analyses. The importance of the Mussio and Smith model is apparent. However, step three of their model is the point of relevance for this discussion: Calculating the job relevance of the examination. The ability to quantify job relevance represents the design of a standard by which satisfactory levels of content representativeness will be judged.

Lawshe's (1975) content validity ratio is an empirical value calculated from independent judgments made by multiple subject matter experts (i.e., incumbents and supervisors) regarding the degree of relevance that a knowledge, skill or ability (KSA) measured by each test item has to job performance. Specifically, each subject matter expert is asked to judge each KSA according to three criteria: (1) the KSA measured by the item is essential to job performance, (2) useful but not essential to job performance, or (3) not necessary

to the performance of the job. The actual content validity ratio (CVR) is then calculated using the following formula:

$$CVR = \frac{(n_e - (N/2))}{N/2}$$

where n_e equals the number of subject matter experts who indicated that that particular item was essential to job performance and where N equals the total number of subject matter experts.

Items are deleted from the examination if the CVR fails to attain statistical significance (see Lawshe, 1975 for the significance table). The average CVR value of the remaining items is computed and then termed the Content Validity Index (CVI). Operationally, the CVI represents the extent to which perceived overlap exists between actual job performance and performance on the examination. This value is the objective standard of the degree of content validity (i.e., content representativeness) for a given instrument.

A field study conducted by Carrier, Dalessio, and Brown (1990) showed a modest relationship between the CVR judgments and empirically determined criterion-related validity, while using an application blank designed to select insurance agents with previous experience. No relationship was found when using an application for selecting agents without previous experience. Nevertheless, the results of this study are encouraging in that they represent the ability to draw similar conclusions from two different validation strategies.

A second method for adding quantification to content validity was based on the behavioral consistency approach (Wernimont & Campbell, 1968) and applied by Schmitt and Ostroff (1986). This approach involves a three-stage process. First, it requires job incumbents to evaluate each job task and KSA on a 7-point scale in terms of (a) the relative amount of time they spend performing this task compared to all other job tasks, (b) the degree of difficulty required to complete this task relative to all other tasks, and (c) the criticality of the task determined by the degree to which improper performance would result in negative consequences. Second, a task importance value is computed for each task (Levine, 1983) according to the following formula: Time Spent + (Difficulty x Criticality). And third, a different set of subordinates then rate the KSAs on three dimensions: (1) the necessity for newly hired employees to possess the KSA (yes or no), (2) the extent to which trouble is likely to occur if the KSA is ignored in the selection system (7-point scale), and (3) the degree to which the KSA is able to distinguish between superior and average workers (7-point scale). Those tasks for which a given KSA was considered to be important then was included as an item on the actual selection test.

The method developed by Lawshe (1975) and the application carried out by Schmitt and Ostroff (1986) signify the ability to produce an empirical estimate of content validity. In and of themselves, those advances are crucial to the psychometric advance of assessing content validity. Ostensibly, the application done by Carrier et at. (1990) represents the first of many projects seeking to define further the correspondence between content and criterion-related validity.

Synthetic Validity. Another successful method of indirect validation is synthetic validity (Balma, 1959; Mossholder & Arvey, 1984; Trattner, 1982). The appeal of synthetic validity is its relevance to those situations where large sample sizes are not readily available. Rather than validating a whole test, or battery of tests, synthetic validity determines the validity of its elemental parts (i.e., subscales, subtests). Once these sections are validated they can then be used individually in other selection system instruments. For example, it would be possible to validate one of the subscales used in the Eligibility Technician examination and then reapplied to other similar job class exams, instead of developing and validating a whole new subscale for the other job classes. An excellent application of this methodology was recently demonstrated by Hollenbeck and Whitener (1988).

Construct Validity. The final validation issue to be discussed is construct validity. It has been said that construct validity is the underlying thread which unifies all validation strategies (Guion, 1977; Tenopyr,

1977). Essentially, the construct validity of a test is the degree to which the test measures a theoretical construct or trait (e.g., leadership, intelligence, mathematical reasoning) based on established interrelationships among behavioral measures. Thus, rather assessing the content area of Comprehending and Applying Eligibility Regulations, construct validity would investigate those dimensions which are more fundamental to that content portion; for example, reading comprehension, planning and organizing, and/or decision-making may be processes which are required to perform that particular subscale.

Empirical investigation of construct validity can be assessed in several ways. However, only three primary techniques will be mentioned briefly. The first method of determination is the multitrait-multimethod paradigm (Campbell & Fiske, 1959). In this situation, traits represent theoretical dimensions (e.g., dominance, achievement motivation, problem analysis) and methods are the different instruments which are used to measure each trait. For the purpose of this report, a job analysis would have to be conducted to determine the relevant job performance dimensions (i.e., traits), and then several selection instruments (e.g., cognitive abilities test, situational interview, and projective test) would have to be designed to measure each dimension. The degree of convergence and psychometric discrimination between the traits would then provide an empirical determination of the selection system's construct validity.

While the multitrait-multimethod process is effective it is also in possession of shortcomings. It is sometimes ambiguous, subjective, and unmanageable when large systems are involved. An improvement over that process is the Analysis of Variance (ANOVA) method (Kavanagh, MacKinney, & Wolins, 1971). This design allows for the partitioning of variance for each of the factors: (1) person variance which indicates the degree of convergence; (2) trait variance which represents discriminant validity; (3) people by methods variance which signifies method variance; and (4) error variance. Clearly, such partitioning of the variance reduces subjectivity, improves accuracy, and the relative magnitude of each relative effect can be estimated. The ANOVA method is also able to conduct analyses on large data sets more effectively than the multitrait-multimethod approach.

Another adaptation of the multitrait-multimethod perspective construct validity is the multitrait-multirater procedure developed by Lawler (1967; see Zedeck & Baker, 1972). Here the different instruments used to assess the constructs are replaced by different raters (e.g., self, peer, supervisory). This procedure is able to fit most employment settings which require that a subordinates performance be evaluated by several assessors. Clearly, this approach has several advantages for establishing criteria and performance appraisal effectiveness.

Additional Issues

Beyond the many issues already discussed, several must be mentioned briefly in that they too deserve additional investigation which will benefit the County. First, the issue of job families (Pearlman, 1980) should be examined in that it may improve the quality of job classification and may serve as a foundation for eventual application of validity generalization (Schmidt & Hunter, 1977). Second, in order to quantify the degree of confidence attained for a selection system further, the principle of cross validation should become a routine procedure (Murphy, 1983, 1984). Third, the effectiveness of the various recruiting sources used by the County should be investigated in psychometric and fiscal terms (Taylor & Schmidt, 1983). Fourth, the degree of dyadic bias present within the performance appraisal process should be assessed (Kraiger & Ford, 1985). And finally, a selection taxonomy should be developed outlining those qualities which are indicative of successful and unsuccessful Eligibility Technicians (R. W. Jacobsen, personal communication, 1990). A taxonomy of this type would facilitate understanding of the criteria used, selection effectiveness, and determination of training needs.

Conclusions and Recommendations

In conclusion, the current written test used to screen applicants for the position of Eligibility Technician is valid and unbiased. The written test was found to be valid up to the point of first-annual performance review. Conversely, the application blank/interview composite was found to possess little

practical use to the selection system. A valid selection instrument has important implications for productivity, cost savings, and legal compliance.

It is important to remember that employee selection should be viewed in terms of an overall human resource system, of which selection is but one portion. The necessity of a sound and rigorous system is evident in the many variables which may affect performance during the tenure of an employee. The following recommendations are made in order to improve and maintain the quality of the human resource system used to hire, develop, motivate, and retain high performing Eligibility Technicians.

1. Delete the use of the application/interview composite from the selection process. Incorporate the use of biographical data, an honesty test, and a situational interview.

2. Implement a multiple criteria model using multiple raters and multiple objective indices of performance.

3. Establish a rater training program for all employees who evaluate work performance.

4. Assess the content validity of the revised selection system annually using one of the quantitative methods described earlier until a suitable sample size is attained.

5. Incorporate a comprehensive investigation of the entire human resource system for Eligibility Technicians: recruitment, selection, training, criterion development, and performance review using a utility analysis orientation.

6. Conduct annual needs assessments to determine the likelihood of implementing a job enrichment/redesign strategy to improve performance and retention.

References

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York, NY: MacMillan Publishing Co., Inc.

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86,* 736-765.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35,* 281-322.

Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). New York, NY: Addison-Wesley.

Ash, R. A., & Levine, E. L. (1985). Job applicant training and work experience evaluation: An empirical comparison of four methods. *Journal of Applied Psychology, 70,* 572-576.

Association Against Discrimination v. City of Bridgeport, 454 F. Suppl. 751 (D. Conn. 1978).

Balma, M. J. (1959). The concept of synthetic validity. *Personnel Psychology, 12,* 395-396.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335-345.

Barrett, G. V., & Kernan, M. C. (1987). Performance appraisal and terminations: A review of court decisions since Brito v. Zia with implications for personnel practices. *Personnel Psychology, 40,* 489-503.

Bernardin, H. J., & Pence, E. C. (1980). Rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60-66.

Blankenship, M. H., & Cesare, S. J. (1990). *Utility analysis of the eligibility technician selection process: The economic impact of test validity* (DHR TR-90-2). San Diego: County of San Diego Department of Human Resources.

Brush, D. H., & Owens, W. A. (1979). Implementation and evaluation of an assessment classification model for manpower utilization. *Personnel Psychology, 32,* 369-383.

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement, 10,* 159-186.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carrier, M. R., Dalessio, A. T., & Brown, S. H. (1990). Correspondence between estimates of content and criterion-related validity values, *Personnel Psychology, 43,* 85-100.

Cascio, W. F. (1987). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cesare, S. J., Blankenship, M. H., Giannetto, P. W., & Mandel, M. Z. (1990). *An assessment of the empirical validity of a selection system for eligibility technicians: A criterion-related approach* (DHR TR-90-3). San Diego: County of San Diego Department of Human Resources.

Civil Rights Act of 1964, Title VII, 42 U. S. C. 2000e.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white student in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dickerson v. U.S. Steel Corp., 472 F. suppl. 1304 (E.D. Penn. 1978).

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. *Personnel Psychology, 41*, 315-327.

EEOC v. Detroit Edison Co., 512 F.2d 301 (6th Cir. 1975).

England, G. W. (1971). *Development and use of weighted application blanks*. Minneapolis, MN: Industrial Relations Center, University of Minnesota.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register, 43*, 38290-38315.

Equal Employment Opportunity Commission (1981). *Guide to pre-employment inquiries*. Washington, D.C.: Equal Employment Opportunity Commission.

Feild, H. S., & Holley, W. H. (1982). The relationship of performance appraisal system characteristics to verdicts in selected employment discrimination cases. *Academy of Management Journal, 25*, 392-406.

Flynn, G. (1990, April 17). Social workers suspected of welfare fraud. *San Diego Union*, pp. 1, 8.

Goldstein, I. L. (1974). The application blank: How accurate are the responses? *Journal of Applied Psychology, 65*, 491-494.

Griggs v. Duke Power Company, 401 U. S. 424 (1971).

Guion, R. M. (1965). *Personnel Testing*. New York: McGraw-Hill Book Company.

Guion, R. M. (1977). Content validity: Three years of talk--what's the action? *Public Personnel Management, 6*, 407-414.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385-398.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*, 250-279.

Harris, W. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42*, 691-726.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Heneman, H. G. (1974). Comparisons of self- and superior ratings of managerial performance. *Journal of Applied Psychology, 59*, 638-642.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811-826.

Hollenbeck, J. R., & Whitener, E. M. (1988). Criterion-related validation for small sample contexts: An integrated approach to synthetic validity. *Journal of Applied Psychology, 73*, 536-544.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*, 349-371.

Imada, A. S. (1982). Social interaction, observation and stereotypes as determinants of differentiation in peer ratings. *Organizational Behavior and Human Performance, 29*, 397-415.

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological Bulletin, 85*, 555-586.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin, 75*, 34-49.

King v. New Hampshire Department of Resources and Economic Development, 15 FEP 669 (1977).

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored ratings scales (BARS) and other rating formats. *Personnel Psychology, 34*, 54-59.

Kleiman, L. S., & Durham, R. L. (1981). Performance appraisal, promotion, and the courts: A critical review. *Personnel Psychology, 34*, 103-121.

Kleiman, L. S., & Faley, R. H. (1978). Assessing content validity: Standards set by the court. *Personnel Psychology, 31*, 701-713.

Kleiman, L. S., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. *Personnel Psychology, 38*, 803-833.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70*, 56-65.

Landy, F. J., & Trumbo, D. A. (1980). *The psychology of work behavior* (2nd ed.). Homewood, IL: Dorsey Press.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies of the situational interview. *Journal of Applied Psychology, 69*, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology, 51*, 369-381.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 565-575.

Lee, R., & Booth, J. M. (1974). A utility analysis of a weighted application blank designed to predict turnover for clerical employees. *Journal of Applied Psychology, 59*, 516-518.

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Predictors, criteria, and significant results. *Personnel Psychology, 24*, 519-533.

Levine, E. L. (1983). *Everything you always wanted to know about job analysis*. Tampa, FL: Mariner Publishing Co.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin, 90*, 125-152.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41*, 283-314.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., Hunter, J. E., Maurer, S., & Russell, J. (1987). *The validity of employment interviews: A review and meta-analysis*. Unpublished manuscript.

Miner, M. G., & Miner, J. B. (1978). *Employee selection within the law*. Washington, D.C.: Bureau of National Affairs, Inc.

Mosel, J. N., & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 44*, 365-369.

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and comparative review. *Journal of Applied Psychology, 69*, 322-333.

Muchinsky, P. M. (1986). Personnel selection methods. In G. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1986* (pp. 37-70). New York: John Wiley and Sons.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology, 36*, 111-118.

Murphy, K. R. (1984). Cost benefit considerations in choosing among cross-validation methods. *Personnel Psychology, 37*, 15-22.

Mussio, S. J., & Smith, M. K. (1973). *Content validity: A procedural manual*. Minneapolis, MN: Civil Service Commission.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology, 41*, 517-535.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609-644). Chicago: Rand-McNally.

Owens, W. A., & Schoenfeldt, L. F. (1983). Towards a classification of persons. *Journal of Applied Psychology, 65*, 569-607.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87*, 1-28.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Rivera v. City of Wichita Falls, 665 F.2d (Fifth Circuit 1982).

Sackett, P. R., & Harris, M. M. (1984). Honesty testing for personnel selection: A review and critique. *Personnel Psychology, 37*, 221-245.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Schmidt, F. L., & Hunter, J. E. (1980). The future of criterion-related validity. *Personnel Psychology, 33*, 41-60

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609-626.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 473-485.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24*, 419-434.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79-101.

Schmitt, N., Gooding, R. Z., Noe, R. D., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39*, 91-108.

Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59*, 583-594.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*, 22-40.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scale. *Journal of Applied Psychology, 47*, 149-155.

Taylor, M. S., & Schmidt, D. W. (1983). A process-oriented investigation of recruitment source effectiveness. *Personnel Psychology, 36*, 343-354.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30*, 47-54.

Thompson, D. E., & Thompson, T. A. (1982). Court standards for job analysis in test validation. *Personnel Psychology, 35*, 865-874.

Trattner, M. H. (1982). Synthetic validity and its application to the Uniform Guidelines validation requirements. *Personnel Psychology, 35*, 383-397.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.

U. S. Department of Labor. (1977). *Dictionary of occupational titles* (4th ed.). Washington, D.C.: U.S. Government Printing Office.

Walls v. Mississippi State Department of Public Welfare, 542 F. Suppl. 281 (N.D. Miss. 1982).

Wards Cove Packing v. Atonio, U.S. Supreme Court 109 S. Ct. 2115 (1989).

Watson v. Fort Worth Bank & Trust, U.S. Supreme Court 108 S. Ct. 2777 (1988).

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72,* 484-487.

Weiner v. County of Oakland, 14 FEP 380 (1976).

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372-376.

Wyer, R. S., & Srull, T. K. (1981). Category accessibility: Some theoretical and empirical issues concerning the processing of social information. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium on personality and social psychology.* Hillsdale, NJ: Lawrence Erlbaum and Associates.

Zedeck, S., & Baker, H. T. (1972). Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance, 72,* 457-466.